Solitons and Collapse in the λ-repressor protein

Andrey Krokhotin,[1,*] Martin Lundgren,[1,†] and Antti J. Niemi[1,2,‡]

[1] *Department of Physics and Astronomy, Uppsala University, P.O. Box 803, S-75108, Uppsala, Sweden*
[2] *Laboratoire de Mathematiques et Physique Theorique CNRS UMR 6083, Fédération Denis Poisson, Université de Tours, Parc de Grandmont, F37200, Tours, France*

The enterobacteria lambda phage is a paradigm temperate bacteriophage. Its lysogenic and lytic life cycles echo competition between the DNA binding λ-repressor (CI) and CRO proteins. Here we scrutinize the structure, stability and folding pathways of the λ-repressor protein, that controls the transition from the lysogenic to the lytic state. We first investigate the super-secondary helix-loop-helix composition of its backbone. We use a discrete Frenet framing to resolve the backbone spectrum in terms of bond and torsion angles. Instead of four, there appears to be seven individual loops. We model the putative loops using an explicit soliton Ansatz. It is based on the standard soliton profile of the continuum nonlinear Schrödinger equation. The accuracy of the Ansatz far exceeds the B-factor fluctuation distance accuracy of the experimentally determined protein configuration. We then investigate the folding pathways and dynamics of the λ-repressor protein. We introduce a coarse-grained energy function to model the backbone in terms of the C_α atoms and the side-chains in terms of the relative orientation of the C_β atoms. We describe the folding dynamics in terms of relaxation dynamics, and find that the folded configuration can be reached from a very generic initial configuration. We conclude that folding is dominated by the temporal ordering of soliton formation. In particular, the third soliton should appear before the first and second. Otherwise, the DNA binding turn does not acquire its correct structure. We confirm the stability of the folded configuration by repeated heating and cooling simulations.

PACS numbers: 05.45.Yv 87.15.Cc 36.20.Ey

I: INTRODUCTION

The transition between the lysogenic and the lytic state in bacteriophage λ infected *E. coli* cell is the paradigm genetic switch mechanism. It is described in numerous molecular biology textbooks and review articles [1]-[7]. The interplay between the lysogeny maintaining λ-repressor (CI) protein and the CRO regulator protein that controls the transition to the lytic state is a simple model for more complex regulatory networks, including those that can lead to cancer in humans.

In the present article we describe the physical properties of the λ-repressor protein, that controls the lysogenic-to-lytic transition. We investigate in detail the stability of its native conformation, the dynamics of the folding process, and the landscape of folding pathways. We find that the folded configuration displays a structure which is unique among all known protein structures. We also conclude that the folding pathways are entirely dominated by the loop regions. In particular, the temporal ordering of loop formation appears to be the decisive factor for the protein's ability to reach its native fold. If solitons form in a wrong order the protein may misfold.

Full crystallographic information of the experimental λ-repressor structure that we use in our investigation is available in Protein Data Bank (PDB) [8] under the code 1LMB. This structure is a homo-dimer with 92 residues in each of the two monomers. It maintains the lysogenic state by binding to DNA with a helix-turn-helix motif that is located between the residue sites 33-51. Throughout this article we shall use the PDB indexing of the residues.

For the statistical analyses that are presented here, we utilize a subset of PDB data that consists of the canonical set of structures with better than 2.0 Å resolution. We have compared the results with the subset that contains only those proteins with better than 2.0 Å resolution and with less than 30% sequence similarity. Our conclusions are independent of the data set, and for illustrative purposes we here use the canonical 2.0 Å set.

This article is composed as follows: We first explain how to describe the geometry of a generic folded protein in terms of its backbone central C_α carbons. We propose that a coarse-grained energy function, that models the backbone geometry, can be constructed with only the C_α coordinates as dynamical variables. We argue that a variant of the discrete non-linear Schrödinger (DNLS) equation is a suitable *Master Equation* to describe folded proteins, in terms of its dark soliton solution. We then proceed to utilize this general framework to study the structure of the λ-repressor protein. We show that the λ-repressor backbone is composed from seven individual soliton solutions of the DNLS equation, within the accuracy of crystallographic structure measurements. In the same vein we propose, that protein folding can be described in terms of a coarse grained model, based on relaxation dynamics. We utilize this to investigate the folding dynamics of the λ-repressor. We conclude that the temporal ordering of soliton formation is important for reaching the correct native state. A wrong ordering in soliton formation can be a cause for misfolding. We observe that the second soliton has a peculiar structure

that sets it apart from any other known structure in all proteins.

II: METHODS:

A: Backbone geometry

Our analysis of the λ-repressor protein will be based on an effective, coarse grained energy function approach that has been recently developed in [9]-[13]. In this approach the protein geometry is described in terms of the backbone C_α atoms. The ensuing bond and torsion angles assume the rôle of the dynamical variables. These angles are constructed as follows: Let $\mathbf{r}_i$ be the coordinate sites of the C_α carbons, where the index $i = 1, ..., N$ runs over all amino acids. For a given protein these coordinates can be read from the PDB. For each site i, we introduce the unit tangent vector

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|} \tag{1}$$

the unit binormal vector

$$\mathbf{b}_i = \frac{\mathbf{t}_{i-1} - \mathbf{t}_i}{|\mathbf{t}_{i-1} - \mathbf{t}_i|} \tag{2}$$

and the unit normal vector

$$\mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i \tag{3}$$

The orthogonal triplet $(\mathbf{n}_i, \mathbf{b}_i, \mathbf{t}_i)$ determines a discrete version of the Frenet frame at each position $\mathbf{r}_i$ of the backbone.

The backbone bond angles are

$$\kappa_i \equiv \kappa_{i+1,i} = \arccos(\mathbf{t}_{i+1} \cdot \mathbf{t}_i) \tag{4}$$

and the backbone torsion angles are

$$\tau_i \equiv \tau_{i+1,i} = \text{sign}\{\mathbf{b}_{i-1} \times \mathbf{b}_i \cdot \mathbf{t}_i\} \cdot \arccos(\mathbf{b}_{i+1} \cdot \mathbf{b}_i) \tag{5}$$

Conversely, if these angles are all known, we can use

$$\begin{pmatrix} \mathbf{n}_{i+1} \\ \mathbf{b}_{i+1} \\ \mathbf{t}_{i+1} \end{pmatrix} = \begin{pmatrix} \cos\kappa\cos\tau & \cos\kappa\sin\tau & -\sin\kappa \\ -\sin\tau & \cos\tau & 0 \\ \sin\kappa\cos\tau & \sin\kappa\sin\tau & \cos\kappa \end{pmatrix}_{i+1,i} \begin{pmatrix} \mathbf{n}_i \\ \mathbf{b}_i \\ \mathbf{t}_i \end{pmatrix} \tag{6}$$

to iteratively construct the frame at position $i + i$ from the frame at position i. Once we have all the frames, we obtain the entire backbone from

$$\mathbf{r}_k = \sum_{i=0}^{k-1} |\mathbf{r}_{i+1} - \mathbf{r}_i| \cdot \mathbf{t}_i \tag{7}$$

Without any loss of generality we may set $\mathbf{r}_0 = 0$, and choose $\mathbf{t}_0$ to point into the direction of the positive z-axis.

We note that the relation (7) does not involve the vectors $\mathbf{n}_i$ and $\mathbf{b}_i$. Consequently we may rotate the $(\mathbf{n}_i, \mathbf{b}_i)$ frame vectors, without affecting the backbone, by selecting an arbitrary linear combination of these two vectors independently at each site i. For this we introduce a local SO(2) transformation that rotates the $(\mathbf{n}_i, \mathbf{b}_i)$ by an angle Δ_i so that the $\mathbf{t}_i$ remain intact,

$$\begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_i \to e^{\Delta_i T^3} \begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_i = \begin{pmatrix} \cos\Delta_i & \sin\Delta_i & 0 \\ -\sin\Delta_i & \cos\Delta_i & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_i \tag{8}$$

where the SO(3) generators are $(T^i)_{jk} = \epsilon_{ijk}$,

$$[T^i, T^j] = \epsilon_{ijk} T^k$$

We combine $\mathbf{n}$ and $\mathbf{b}$ into the complex vector

$$\mathbf{n} + i\mathbf{b}$$

and rewrite (8) as

$$\mathbf{n}_i + i\mathbf{b}_i \to e^{i\Delta_i}(\mathbf{n}_i + i\mathbf{b}_i) \equiv \mathbf{e}_i^1 + i\mathbf{e}_i^2 \tag{9}$$

The frame rotation (8) corresponds to the following transformation in the bond and torsion angles,

$$\kappa_i \, T^2 \;\to\; e^{\Delta_i T^3}(\kappa_i T^2)\, e^{-\Delta_i T^3} \tag{10}$$

$$\tau_i \;\to\; \tau_i + \Delta_{i-1} - \Delta_i \tag{11}$$

Since the transformation (10), (11) leaves $\mathbf{t}_i$ intact it has no effect on the backbone.

A priori, the fundamental range of the bond angle κ_i is $\kappa_i \in [0, \pi]$. For the torsion angle the range is $\tau_i \in [-\pi, \pi)$. Consequently we may identify (κ_i, τ_i) with the canonical latitude and longitude angles of a two-sphere $\mathbb{S}^2$. However, in the sequel we find it useful to extend the range of κ_i into $[-\pi, \pi]$ $mod(2\pi)$, but with no change in the range of τ_i. We compensate for this two-fold covering of $\mathbb{S}^2$ by introducing the following discrete $\mathbb{Z}_2$ symmetry

$$\begin{array}{lll} \kappa_k \;\to\; -\,\kappa_k & & \text{for all } \; k \geq i \\ \tau_i \;\to\; \tau_i - \pi & & \end{array} \tag{12}$$

We note that this is a special case of (10), (11), with

$$\begin{array}{ll} \Delta_k = \pi & \text{for } \; k \geq i+1 \\ \Delta_k = 0 & \text{for } \; k < i+1 \end{array}$$

The regular protein secondary structures correspond to definite values of (κ_i, τ_i). For example standard α-helix is

$$\alpha - \text{helix}: \quad \begin{cases} \kappa \approx \frac{\pi}{2} \\ \tau \approx 1 \end{cases} \tag{13}$$

and standard β-strand is

$$\beta - \text{strand}: \quad \begin{cases} \kappa \approx 1 \\ \tau \approx \pi \end{cases} \tag{14}$$

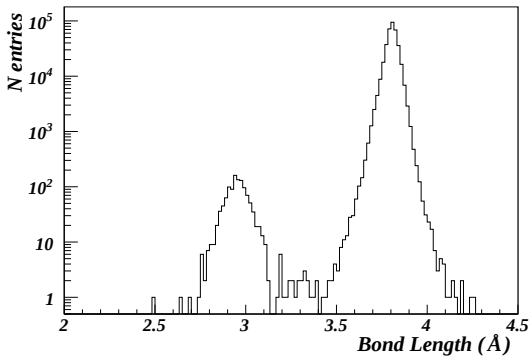

FIG. 1: The distribution of the C_α-C_α bond length in PDB. The fluctuations around the average value $d \approx 3.8$ Å are small, the secondary peak around 2.9 Å is due to *cis*-proline. Note logarithmic scale.

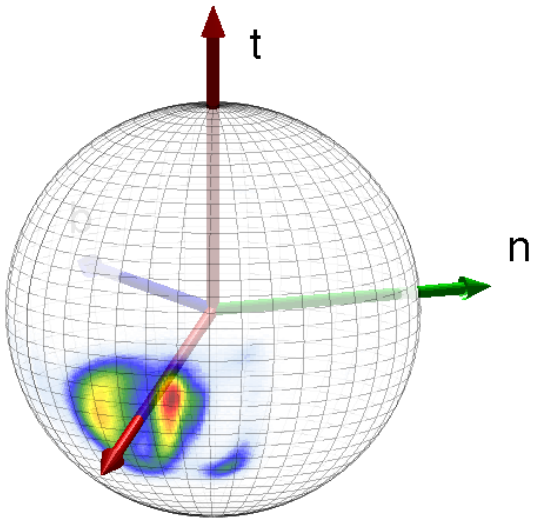

FIG. 2: (Color online) In the Frenet frames the fluctuations in the direction of the vectors $\mathbf{u}_i$ around their average value, given by (18), (20) and denoted by the (red) lateral vector, are relatively small. The left-hand side of the horse-shoe corresponds to β-strands, the right-hand side to α-helices and the connecting region at the bottom corresponds to loops. The minuscule isolated island corresponds to the "left-handed α" region.

Similarly we can describe all the other regular secondary structures such as 3/10 helices, left-handed helices *etc.* with definite constant values of κ_i and τ_i. Loops are configurations that interpolate between these regular structures, so that along a loop the values of (κ_i, τ_i) are variable.

Finally, we compute the average value of the bond length in (7) using PDB. The result shown in Figure 1 is constructed using the canonical set of protein structures which are measured with better than 2.0 Å resolution. The average value is concentrated around

$$|\mathbf{r}_{i+1} - \mathbf{r}_i| = d \approx 3.8\,\text{Å} \tag{15}$$

In our theoretical analysis we use the fixed bond length value (15). We also impose the forbidden volume (steric) constraint

$$|\mathbf{r}_i - \mathbf{r}_k| \geq 3.8\ \text{Å} \quad \text{for} \quad |i-k| \geq 2 \tag{16}$$

between the backbone C_α atoms. This condition is well respected by folded protein structures in PDB.

B: Side-chain geometry

Following [13] we characterize the side-chain directions in terms of directional vectors $\mathbf{u}_i$ that point from the C_α towards the corresponding C_β carbon. At each C_α

$$\mathbf{u}_i = \begin{pmatrix} \sin\theta_i \cdot \cos\varphi_i \\ \sin\theta_i \cdot \sin\varphi_i \\ \cos\theta_i \end{pmatrix} \tag{17}$$

The latitude angle θ_i counts deviation from the direction of the corresponding Frenet frame tangent vector $\mathbf{t}_i$. When $\theta_i = 0$ the $\mathbf{t}_i$ and $\mathbf{u}_i$ are parallel. Note that the angle θ_i remains invariant under the rotation (8). We can compute the values of θ_i from the PDB. As shown in Figure 1 the range of variations in θ_i are quite small, it fluctuates around

$$< \theta > \approx\ 1.98\ \text{(rad)} \tag{18}$$

The longitude φ_i in (17) measures distance from the direction of the Frenet frame normal vector $\mathbf{n}_i$. It is the azimuthal angle between $\mathbf{n}_i$ and the projection of $\mathbf{u}_i$ on the normal plane spanned by $(\mathbf{n}_i, \mathbf{b}_i)$. Under the frame rotation (8) we have

$$\varphi_i \ \rightarrow\ \varphi_i + \Delta_i \tag{19}$$

and consequently the values of φ_i depend on the framing. From PDB we find that in the Frenet frames the values of φ_i are subject to relatively small fluctuations around the average value

$$< \varphi > \approx\ -2.43\ \text{(rad)} \tag{20}$$

As shown in Figure 2, it is remarkable that the direction of $\mathbf{u}_i$ nutates in a very regular horse-shoe shaped manner, that reflects the underlying secondary structure environment. This proposes that there is a strong local coupling between the two angular variables θ_i and φ_i, that depends on the secondary structure.

The notable expection from (18), (20) is the left-handed loop region. It is visible in Figure 2 as a mi-

nuscule isolated region, with

$$\begin{array}{c} <\theta> \approx \ 2.25 \ \text{(rad)} \\ <\varphi> \approx \ -1.90 \ \text{(rad)} \end{array} \tag{21}$$

But this is also *quite* close to the values in (18), (20).

C: Backbone energy and solitons

Proteins display a hierarchy which is determined by the spatial length scale. As the length scale increases, shorter distance dynamical variables become gradually disengaged. Following the general concept of universality [14]-[17], we utilize this hierarchy of scales to systematically coarse grain the energy function. At each level of hierarchy we retain explicitly only those variables that remain relevant. The variables that are less and less so as the distance scale increases are accounted for effectively, through the functional form and the values of parameters in the various individual energy contributions that involve the remaining relevant variables only.

In [9]-[13], see also [18], it has been argued that a Landau free energy function that aims to compute the overall fold geometry can be based solely on those variables that determine the positions of the central C_α atoms. Since the fluctuations in the bond lengths are minimal, see Figure 1, the leading order contribution to the energy then involves only the bond and torsion angles for the C_α backbone. The functional form of the energy can be uniquely determined by symmetry considerations. For this we note that *any* backbone energy function that involves only the bond and torsion angles must remain invariant under the SO(2) transformation (10), (11). Indeed, previously it has already been shown [9]-[13] how this SO(2) gauge invariance allows us to uniquely deduct the functional form of the energy function, in the long distance limit where any higher order non-local contribution can be ignored. In this limit the energy function can only involve the following terms [9]-[11],

$$E = -\sum_{i=1}^{N-1} 2\,\kappa_{i+1}\kappa_i + \sum_{i=1}^{N}\Big\{ 2\kappa_i^2 + q\cdot(\kappa_i^2 - m^2)^2$$

$$+\frac{d_\tau}{2}\,\kappa_i^2\tau_i^2 - b_\tau\kappa_i^2\tau_i - a_\tau\tau_i + \frac{c_\tau}{2}\tau_i^2\Big\} \tag{22}$$

The detailed derivation of (22) is presented in [9]. It involves the introduction of frame (gauge) invariant combinations of the Frenet frame bond and torsion angles [18], [19]. For the present purposes it suffices to observe, that (22) coincides with the one dimensional discretized Abelian Higgs Model Hamiltonian in the unitary gauge, in terms of the Frenet frame bond and torsion variables.

We can also recognize (22) as a variant of the discrete nonlinear Schrödinger (DNLS) equation [20]: The first sum together with the three first terms in the second sum comprise exactly the energy of the standard DNLS equation when expressed in terms of the Hasimoto variable of fluid mechanics [11], [20]. The fourth (b_τ) is a conserved quantity in the DNLS hierarchy, the "momentum", and the fifth term (a_τ) is the conserved "helicity". The last (c_τ) term is the (non-conserved) Proca mass term that we include for completeness.

The energy function (22) does not purport to explain the fine details of the atomary level mechanisms that give rise to protein folding. Rather, it examines the properties of a folded protein backbone in terms of universal physical arguments along the lines of [14]-[17]. Indeed, the functional form (22) is deeply anchored in the elegant mathematical structure of integrable hierarchies [20]. Within this framework no term beyond those in the integrable hierarchy can be added without violating the underlying general and elegant mathematical principles. In this sense, (22) is the *universal* long distance limit that would emerge from *any* microscopic level Schrödinger operator, whenever we truncate the Landau free energy to explicitly include only the backbone bond and torsion angles.

Remarkably, we have found that despite the very general nature of argumentation that leads us to adopt the energy function (22), it is fully capable of describing folded protein backbones with sub-atomic precision of around 0.5 Å, and even better [21]. This is due to the observation [10], [11], that (22) describes proteins in terms of solitons that are the paradigm structural self-organizers. Indeed, solitons are tremendously robust in their ability to preserve the form under both quantum mechanical and thermal fluctuations.

We derive the relevant soliton profile as follow: We first introduce the τ-equation of motion

$$\frac{\partial E}{\partial \tau_i} = d_\tau\kappa_i^2\tau_i - b_\tau\kappa_i^2 - a_\tau + c_\tau\tau_i = 0$$

from which we solve

$$\tau_i[\kappa] = \frac{a_\tau + b_\tau\kappa_i^2}{c_\tau + d_\tau\kappa_i^2} \tag{23}$$

Even though there are four parameters in (23) one of them, the overall scale, cancels out. In the sequel we shall choose $a_\tau = -1.0$ so that for an α-helix (13) we have

$$\tau_i[\alpha] = \frac{1 + b_\tau\kappa_i^2}{c_\tau + d_\tau\kappa_i^2} \approx 1 \quad \mod(2\pi)$$

and for a β-strand (14)

$$\tau_i[\beta] = \frac{1 + b_\tau\kappa_i^2}{c_\tau + d_\tau\kappa_i^2} \approx \pi \quad \mod(2\pi)$$

When we use (23) to eliminate the torsion angles we

get for the bond angles the energy

$$E[\kappa] = -\sum_{i=1}^{N-1} 2\,\kappa_{i+1}\kappa_i + \sum_{i=1}^{N} 2\kappa_i^2 + V[\kappa_i] \qquad (24)$$

where

$$V[\kappa] = -\left(\frac{b_\tau c_\tau - a_\tau d_\tau}{d_\tau}\right)\cdot\frac{1}{c_\tau + d_\tau\kappa^2}$$

$$-\left(\frac{b_\tau^2 + 8qm^2}{2b_\tau}\right)\cdot\kappa^2 + q\cdot\kappa^4 \qquad (25)$$

The functional form of (25) is familiar from various studies in mathematical physics: The first term relates to the potential energy in a Calogero-Moser system. The second and third terms have the conventional form of a symmetry breaking double-well potential. Depending on the parameter values we may be either in the broken symmetry phase where κ and τ both acquire a non-vanishing and constant ground state value, or in the symmetric phase where κ vanishes.

In applications to proteins, regular structures such as helices (13) and strands (14) correspond to different broken symmetry ground states of the energy. Furthermore, the numerical value of the first term in (25) is often small in comparison to the second and third, and so is b_τ^2 in comparison to $8qm^2$ so that for an α-helix (13)

$$m \approx \frac{\pi}{2} \qquad (26)$$

and for a β-strand

$$m \approx 1.0 \qquad (27)$$

In [10], [11] it has been shown that loops, which are regions where (κ_i, τ_i) are variable, can be constructed in terms of the dark soliton solution of the generalized discrete nonlinear Schrödinger equation that derives from the energy (24),

$$\kappa_{i+1} = 2\kappa_i - \kappa_{i-1} + \frac{dV[\kappa]}{d\kappa_i^2}\kappa_i \qquad (i = 1, ..., N) \qquad (28)$$

where we set $\kappa_0 = \kappa_{N+1} = 0$. This is the *Master equation* that we use to compute the shape of a folded protein C_α backbone.

D: Soliton Ansatz

We do not know the explicit form of the dark soliton solution to the present discrete nonlinear Schrödinger equation. But a numerical approximation can be easily constructed using the procedure described in [11]. Furthermore, since it turns out that in the case of proteins the first term in (25) is small, an excellent approximation [12] is given by the *naive* discretization of the continuum dark NLSE soliton [20],

$$\kappa_i = \frac{(\mu_1 + 2\pi N_1)\cdot e^{\sigma_1(i-s)} - (\mu_2 + 2\pi N_2)\cdot e^{-\sigma_2(i-s)}}{e^{\sigma_1(i-s)} + e^{-\sigma_2(i-s)}} \qquad (29)$$

Here s is a parameter that determines the backbone site location of the soliton center. This is the center of the fundamental loop that we describe by the soliton. The $\mu_{1,2} \in [0, \pi]$ are parameters. In the case of proteins the values of $\mu_{1,2}$ are entirely determined by the adjacent helices and strands. Far away from the soliton center we have

$$\kappa_i \;\to \begin{cases} \mu_1 & \mod(2\pi) \quad i > s \\ -\mu_2 & \mod(2\pi) \quad i < s \end{cases}$$

For α-helices and β-strands the $\mu_{1,2}$ values are determined by (13), (14). Negative values of κ_i are related to the positive values by (12).

The N_1 and N_2 constitute the integer parts of $\mu_{1,2}$ and for simplicity we shall take $N_1 = N_2 \equiv N$. This integer is like a covering number, it determines how many times κ_i covers the fundamental domain $[0, \pi]$ when we traverse the soliton once. We introduce this integer for the following reason: The Ansatz (29) is monotonic but in general the values of $\kappa_i \in [0, \pi]$ that we obtain from PDB are not. Whenever we encounter a site i where κ_i in the PDB data fails to be monotonic, we either add or subtract 2π to its value, to regain a monotonic data profile. This shift does not have any effect on the backbone geometry. In this manner we utilize the multi-valuedness of κ_i to convert any sequence $\{\kappa_i\}$ into a monotonic one, that we can then approximate by the Ansatz (29).

Note that for $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$ we recover the hyperbolic tangent. In this case the two regular secondary structures before and after the loop are the same. Moreover, *only* the (positive) σ_1 and σ_2 are intrinsically loop specific parameters. They specify the length of the loop and as in the case of the $\mu_{1,2}$ they are combinations of the parameters in (22).

Similarly, in the case of the torsion angle there is only one loop specific parameter in (23). The overall, common scale of the four parameters is again irrelevant and two of the remaining three parameters are fixed by the regular secondary structures that are adjacent to the loop.

Long protein loops, and entire super-secondary protein structures such as a helix-loop-helix can be constructed by combining together solitons (28), (23). A typical short super-secondary structure that is described by a single soliton involves at least 15-20 different amino acids, often even more. As a consequence our DNLS equations and the explicit soliton Ansatz compute the 45-60 spatial coordinates of the ensuing C_α carbons in terms of of the five essential universal parameters in (22). This implies that the DNLS equations comprises a highly underdetermined system of equations, and the key physical principles of our approach are experimentally testable.

In [21] it has been shown using the Ansatz (29) that over 92% of PDB configurations can be constructed in terms of 200 explicit soliton profiles. This makes a strong case that the solitons of the DNLS equation are the modular building blocks of folded proteins [21].

E: Side-chain energy function

In order to account for the C_β contribution to the protein free energy, we augment (22) by terms that involve the variables (θ_i, φ_i) in (17). We shall assume that side-chain directions are *locally slaved* to the backbone. By an explicit analysis of PDB structures using Frenet frames this can be confirmed to be the case [22], as shown in Figure 2.

The C_β latitude angles θ_i are gauge *i.e.* frame invariant: The θ_i are entirely determined by the tangent vectors $\mathbf{t}_i$, and consequently can not depend on the choice of framing. To the leading order we may then assume that each θ_i interacts locally, with the corresponding κ_i only. The leading order contribution is obtained by Taylor expanding a general interaction potential around the (κ_i dependent) equilibrium values of the θ_i,

$$E_\theta = \sum_{i=1}^{N} \left\{ \frac{d_\theta}{2}\, \kappa_i^2 \theta_i^2 - b_\theta \kappa_i^2 \theta_i - a_\theta \theta_i + \frac{c_\theta}{2} \theta_i^2 \right\} + \dots \quad (30)$$

where the additional terms are of higher order in powers of κ_i and θ_i. From this we solve for θ_i,

$$\theta_i = \frac{a_\theta + b_\theta \kappa_i^2}{c_\theta + d_\theta \kappa_i^2} \quad (31)$$

Again, as in the case of (23), the overall scale cancels which leaves us with only three independent parameters. As visible from Figure 2, the fluctuations in θ_i around the average value (18) are small. From this Figure we also learn [22] that these fluctuations are slaved to the backbone geometry, which is dictated by the κ_i. This confirms that the present approximation (30) is reasonable.

Unlike θ_i, the C_β longitude angle φ_i does not remain intact under the frame rotation (8) but transforms according to (19). Consequently it can form a SO(2) frame (*i.e.* gauge) invariant combination with the backbone torsion angle (5). Two examples of gauge invariant combinations are

$$\tau_i - \varphi_{i-1} + \varphi_i$$

and

$$\varphi_i + \sum_{k=1}^{i} \tau_k \quad (32)$$

We prefer to proceed with the second one, it is local in φ_i. (The first is a difference of the second.) As in (22) we specify the unitary gauge, which amounts to selecting the Frenet framing along the backbone.

As visible in Figure 2, the fluctuations in φ_i are about as small as those in θ_i. Moreover, in the combinations (32) the torsion angles τ_i are determined locally by the bond angles according to (23). Consequently we may also Taylor expand the φ_i contribution to the energy, and following (30) we conclude that the leading order contribution is of the form

$$E_\varphi = \sum_{i=1}^{N} \left\{ \frac{d_\varphi}{2}\, \kappa_i^2 \varphi_i^2 - b_\varphi \kappa_i^2 \varphi_i - a_\varphi \varphi_i + \frac{c_\varphi}{2} \varphi_i^2 \right\} + \dots \quad (33)$$

This slaves φ_i to the backbone κ_i according to

$$\varphi_i = \frac{a_\varphi + b_\varphi \kappa_i^2}{c_\varphi + d_\varphi \kappa_i^2} \quad (34)$$

Again only three of the four parameters are independent, the overall scale drops out.

F: Total energy

We combine (22), (30) and (33) to arrive at the total energy

$$E = E_\kappa + E_\tau + E_\theta + E_\varphi \quad (35)$$

$$= -\sum_{i=1}^{N-1} 2\, \kappa_{i+1} \kappa_i + \sum_{i=1}^{N} \left\{ 2\kappa_i^2 + q \cdot (\kappa_i^2 - m^2)^2 \right\} \quad (36)$$

$$+ \sum_{i=1}^{N} \left\{ \frac{d_\tau}{2}\, \kappa_i^2 \tau_i^2 - b_\tau \kappa_i^2 \tau_i - a_\tau \tau_i + \frac{c_\tau}{2} \tau_i^2 \right\} \quad (37)$$

$$+ \sum_{i=1}^{N} \left\{ \frac{d_\theta}{2}\, \kappa_i^2 \theta_i^2 - b_\theta \kappa_i^2 \theta_i - a_\theta \theta_i + \frac{c_\theta}{2} \theta_i^2 \right\} \quad (38)$$

$$+ \sum_{i=1}^{N} \left\{ \frac{d_\varphi}{2}\, \kappa_i^2 \varphi_i^2 - b_\varphi \kappa_i^2 \varphi_i - a_\varphi \varphi_i + \frac{c_\varphi}{2} \varphi_i^2 \right\} + \dots \quad (39)$$

Since the variations in (θ_i, φ_i) are much smaller than those in τ_i, the ensuing contributions E_θ and E_φ are also much smaller than E_τ. Furthermore, according to (39) the backbone torsion angles τ_i and the side-chain angles (θ_i, φ_i) are all slaved to the backbone bond angles κ_i. As a consequence the DNLS equation (28) is the *Master Equation* that entirely determines the geometry of a folded protein: The C_α–C_β backbone is constructed by first solving for κ_i. The remaining two angles (θ_i, φ_i) are then constructed in terms of the κ_i using (23), (31) and (34).

G: Parameters

The energy function (35)-(39) involves a number of parameters. Eventually, we would like to compute their numerical values directly from the amino acid sequence. But at the moment this has not yet been achieved.

A priori it appears that the number of parameters needed in (35)-(39) to describe an entire protein backbone, could be quite large. However, due to the presence of the dark soliton the number of parameters is actually remarkably small: For each super-secondary structure such as a helix-loop-helix, whenever the loop can be described in terms of a single soliton solution to (28), the potential (25) has only four independent parameter combinations. In addition of q and m that characterize the second and third term, there are only two essential parameters in the first term. Three of these four parameters can be given the following interpretations. Two of them determine the values of κ_i in the ground states such as (13), (14) that are located along the backbone before and after the soliton *i.e.* the type of the helix that precedes and follows the soliton. The third parameter determines the length of the loop. The fourth parameter can then be included as one of the three independent parameters in the torsion profile (23). It can be attributed to the length of the soliton, in terms of torsion. In addition, in the soliton profile of κ_i there is the parameter that specifies the position of the soliton along the backbone. This is an additional parameter that emerges from the periodicity of the lattice structure (lattice translation invariance).

Since the overall scale in (23) cancels out, the two remaining parameter combinations in addition of the loop length, become determined by the values of τ_i in the ground states surrounding the soliton *i.e.* the type of the helix as in (13), (14).

In this way we arrive at the conclusion that for the backbone, the only loop specific parameters are those that determine the lengths of the solitons. All additional parameters in the energy function determine the regular secondary structures such as (13) and (14). The profiles of all loops are completely fixed by the *unique* dark solution solution to (28).

Similarly, we conclude from (31), (34) that in the equations that determine the Frenet frame orientations of the C_β carbons, there is only one loop specific parameter in both θ_i and φ_i. In each equation, the overall scale factor cancels out and the values of the additional two independent parameters are fully specified by the regular secondary structures adjacent to the loop.

Since (35)-(39) aims to predict the 45-60 space coordinate of C_α, and the corresponding 30-45 directional coordinates of the C_β in terms of 11 essential parameters, we have a highly underdetermined system of equations. This implies that the physical principles of our approach are experimentally testable.

In [21] we have found that most crystallographic protein structures in PDB can be described in a modular fashion and with experimental B-factor precision, by combining together no more than 200 explicit soliton profiles. We propose that by learning how to compute the parameter values directly from the sequence, the geometric shape of most folded proteins can be constructed simply by solving the Master equation (28).

H: Fluctuations around solitons

As such, the equations (28), (23), (31), (34) describe the critical points of the energy function (35). This energy function should be duly interpreted as describing the effective long distance limit of the full microscopic second-quantized Schrödinger operator. As such, (35) then relates to proteins in the sense of a semi-classical approach. This kind of semi-classical description is common in quantum field theories. There, it is often boldly used to describe phenomena at length scales that are several orders of magnitude smaller than anything which may have any direct relevance to proteins. We now wish to estimate the short distance scale, at which we expect the semiclassical approach to proteins based on (35), to break down due to quantum mechanical zero-point fluctuations.

The backbone profile (28), (23) describes the C_α lattice in the limit where thermal fluctuations vanish. But even near zero temperature the protein remains subject to residual zero-point fluctuations that can not always be ignored. It is difficult to estimate and even harder to accurately calculate the amplitude of these zero-point fluctuations. For a realistic order of magnitude estimate we inspect the distribution of the B-factors that characterize experimental uncertainties in PDB data. We use all those PDB structures where the crystallographic measurements have been made at temperatures less than 50K. The result is displayed in Figure 3. It shows that for the C_α carbons the zero point fluctuations have a lower bound which is in the vicinity of 0.15 Å. Consequently we estimate that the precision of our semi-classical approach can at best be of the same order of magnitude. We account for these zero point fluctuations by dressing our (semi)classical backbone profiles with a tubular dominion that has a radius of 0.15 Å. In particular, this tubular dominion accounts for the bond length fluctuations, around the average distance (15) between the neighboring C_α carbons. As can be seen from Figure 1, the fluctuations in the $C_\alpha - C_\alpha$ distances are normally within range of 0.15 Å.

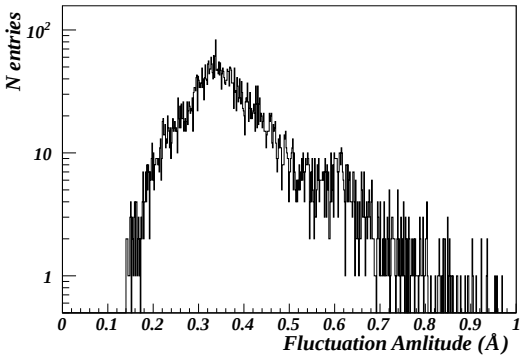

FIG. 3: The number of entries in PDB with temperature below 50K *vs.* Debye-Waller fluctuation distance. Note the logarithmic scale.

III: λ-REPRESSOR PROTEIN AS A MULTISOLITON CONFIGURATION

A: Loop spectrum of 1LMB

We start our soliton-based investigation of the λ-repressor protein by analyzing its (κ_i, τ_i) spectrum. This will identify the putative soliton content. We use the first chain of the homo-dimer with the PDB code 1LMB. The structure is conventionally interpreted as a four loop configuration, and the second loop is the DNA binding one.

From the PDB file we read the C_α coordinates. We compute the tangent vectors from (1) and the binormal vectors from (2), and the bond and torsion angles from (4) and (5). We construct these angles using the standard convention that $\kappa_i \in [0, \pi]$. We locate the regions where the torsion angles τ_i are subject to large fluctuations. In these regions we judiciously implement the transformation (12). This identifies the soliton structures in the loops. Both the motivation and the technical details of the soliton identification procedure are described in [10] and [13].

In the left hand side of Figure 4 we show the (κ_i, τ_i) spectra that we obtain from the PDB data using the standard differential geometric convention that curvature is positive $\kappa_i > 0$. Each of the four figures describes the spectra over one of the four loops of 1LMB, as they are identified in PDB. We observe that in each Figure 4 on the left hand side, there is a region where the torsion angle τ_i fluctuates rapidly between positive values and negative values. On general grounds [10], [13] a region where the torsion angle is solely positive and only subject to small variations is a putative regular secondary structure. On the other hand, a region where the torsion angle fluctuates between positive and negative values is an indicative of an inflection point [13], this kind of fluctuation

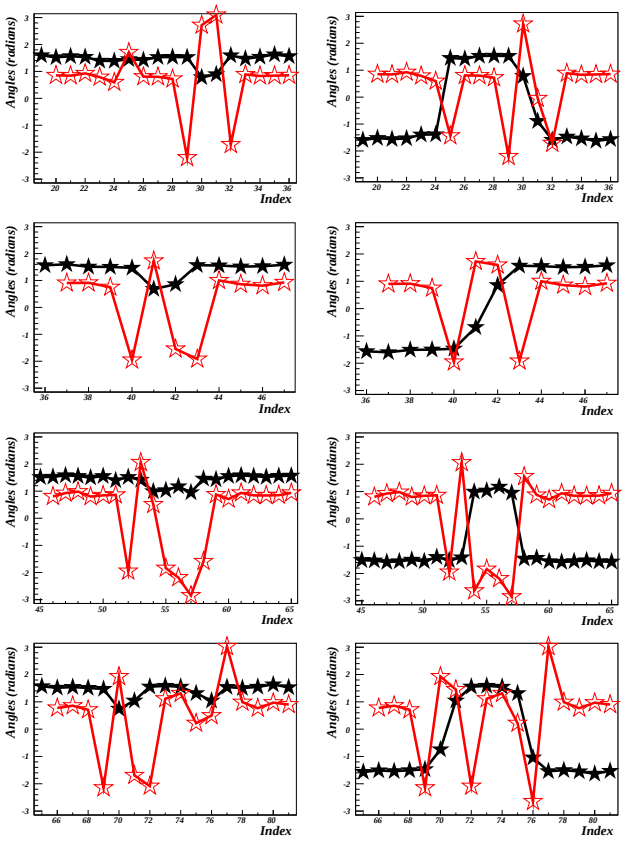

FIG. 4: (Color online) On the left column, we show the (κ_i, τ_i) spectrum for the four PDB loop structures of 1LMB, that we compute with the prevailing differential geometric convention that curvature is a nonnegative quantity. Black is the bond angle κ_i, grey (red) is the torsion angle τ_i. On the right we display the corresponding spectra after we have identified the soliton profiles in κ_i, using (12). All PDB loop structures except for the second, display the profile of a soliton pair. Only the second PDB loop can be identified as a single soliton state.

suggests the presence of solitons. By judiciously applying the gauge transformation (12) in the regions where τ_i fluctuates between positive and negative values, we find that the κ_i profiles in each of the left hand side Figure 4 indeed describe solitons: Comparing with (13), (14) we conclude that the first loop structure in the PDB data is a pair of solitons separated by a short α-helix. The second PDB loop is a single soliton that interpolates between two α helices. The third loop structure in the PDB data is a pair of solitons separated by a short β-strand. Finally, the fourth PDB loop is a pair of solitons, separated by a short α-helix.

Notice that our spectral analysis based structure identification is a refinement of the conventional one, which utilizes techniques such as the presence or absence of hydrogen bonds to conclude whether a site is part of a regular secondary structure or part of a loop. Consequently

very short helical structures that become clearly visible in our (κ_i, τ_i) spectral analysis, can be interpreted differently in more conventional approaches.

B: Soliton Ansatz

We proceed to evaluate the parameters in (29), (23) that give our best fit to those seven soliton profiles, identified by analyzing the (κ_i, τ_i) spectrum. In Table 1 we present our best parameter values. The parameters can be computed by various different techniques. Here we have chosen a Monte Carlo search that is fast and gives very good accuracy. In Table 1 we also identify the corresponding super-secondary structures by their PDB backbone indices. Note that since our approach is based on the spectral analysis of bond and torsion angles and since the definition of a bond angle involves three sites while that of the torsion angle involves four, three residue indices at both ends of the 1LMB chain are absent in the (κ_i, τ_i) list. Notice also that we have introduced some overlap between the different super-secondary structures. This ensures that we can eventually combine them together into the full 1LMB backbone. Moreover, in the case of all except the fourth soliton, we have utilized the multi-valuedness in the definition of the bond angle to extend the range of its values outside of the fundamental domain $\kappa_i \in [0, \pi]$. This corresponds to selecting non-vanishing integers N_1, N_2 in the Ansatz (29). For simplicity we have limited our Monte Carlo search to the symmetric case $N_1 = N_2$, but for better accuracy the soliton profiles could also be given asymmetric integer parts. The numerical values of the bond angles κ_i that we compute from (29) using the parameter values in Table I, are to be reduced onto the fundamental domain $[0, \pi]$ using the $mod(2\pi)$ multivaluedness.

TABLE I: Our best parameter values for each of the seven solitons in Figure 2. Note that m_1, m_2 are determined mod (2π).

Soliton	c_1	c_2	m_1	m_2
(8,30)	2.00441	1.99595	26.65124	26.68412
(26,39)	2.94889	2.95201	70.67882	70.60369
(36,47)	2.89729	2.90755	39.27387	39.22546
(46,59)	2.97927	3.00015	1.07948	1.52942
(56,65)	2.96486	2.97087	26.69087	26.25280
(62,74)	2.94948	2.94547	20.43071	20.38220
(74,87)	2.89725	2.89945	89.50870	89.55252

Soliton	s	a/b	d/b
(8,30)	24.50259	$-9.9921 \cdot 10^{-2}$	$4.2191 \cdot 10^{-5}$
(26,39)	30.49642	$-1.5114 \cdot 10^{-7}$	$1.0662 \cdot 10^{-11}$
(36,47)	41.39325	$-5.3794 \cdot 10^{-7}$	$7.4566 \cdot 10^{-11}$
(46,59)	53.67225	$5.1477 \cdot 10^{-7}$	$-5.1529 \cdot 10^{-7}$
(56,65)	57.85123	$-9.62942 \cdot 10^{-8}$	$1.45097 \cdot 10^{-12}$
(62,74)	70.22069	$-9.27151 \cdot 10^{-7}$	$3.05202 \cdot 10^{-10}$
(74,87)	75.56315	$-7.13705 \cdot 10^{-7}$	$1.8457 \cdot 10^{-11}$

We recall from Section II D, that the introduction of non-vanishing values of N_1 and N_2 enables us to account for the non-monotonic profile that the bond angles of the PDB configuration display when restricted into the fundamental domain: At each point where the profile of the PDB data becomes non-monotonic, we simply add (or subtract) 2π until we obtain a monotonic profile. In this manner, by allowing the bond angle to take values over a larger domain, the κ_i profile of the PDB data over each soliton can be made monotonic which improves the accuracy of the fitted soliton profile (29). See Figure 5 where we display the second PDB loop together with its approximation by the Ansatz (29) in the extended range, as an example.

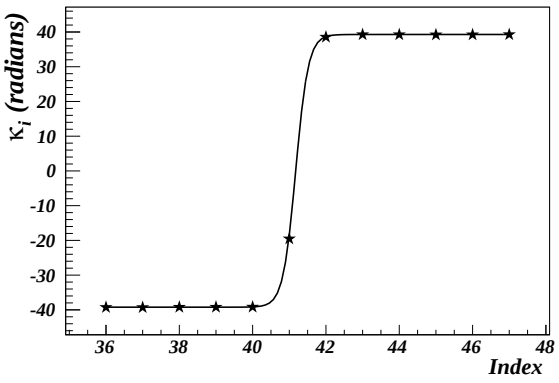

FIG. 5: The data points correspond to the second loop of 1LMB, after we have properly extended the range of the bond angles, by using the inherent multivaluedness of these angles. The interpolating function describes the Ansatz (29) with the parameter values given in Table I.

We compute the torsion angles τ_i from (23), before implementing the 2π reduction of the bond angles. We then reduce the ensuing values of the τ_i into the fundamental domain $\tau_i \in (-\pi, \pi]$ using the 2π periodicity. The underlying multivaluedness entirely accounts for the fluctuations in the τ_i profile.

In Figure 6 we compare the explicit backbone profiles that we have computed from (29), (23), (6), (7) using the parameter values in Table I, with the 1LMB data in PDB. We find that for each soliton, our backbone profiles describe the structural motifs of 1LMB with a precision that is *substantially* better than the experimental precision which is determined by B-factors. This persist even when we account for the 0.15 Å estimate of the zero point fluctuations around our solitons. With these highly precise soliton profiles we can *unambiguously* con-

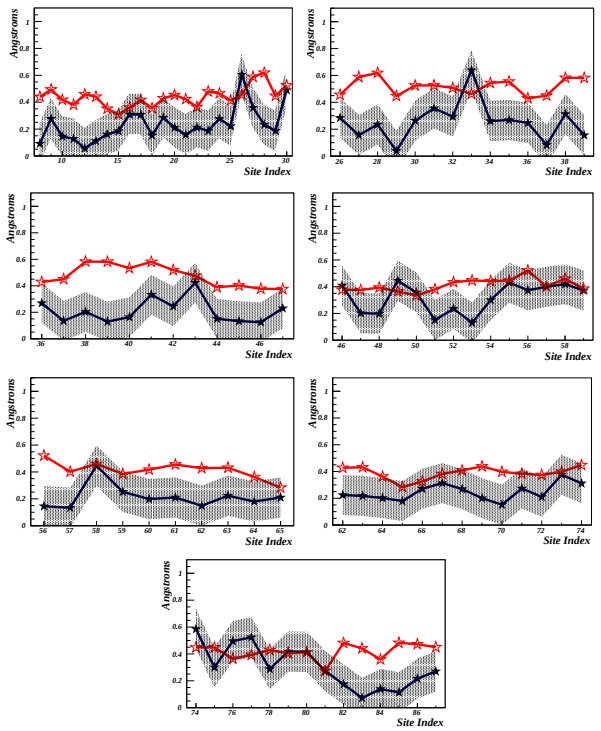

FIG. 6: (Color online) The distance between the PDB backbone of the first 1LMB chain and its approximation by the seven solitons (29), (23) as a function of the residue number. The black line denotes the distance between the soliton and the corresponding PDB configuration, the grey area around the black line describes the estimated 0.15 Å zero point fluctuation distance around solitons, obtained from Figure 3. The grey (red) line denotes the Debye-Waller distance that is computed form the B-factors in PDB.

clude that 1LMB has a total of seven solitons. Two of the α-helices and the sole β-strand are so short that the ensuing soliton pairs are interpreted as single loops in the conventional, hydrogen bond based analysis of the PDB data. The only motif where our construction identifies a single PDB loop as a single isolated soliton, is the DNA binding one. All of the remaining three loops consist of a pair of solitons with profile (29), (23) that are separated from each other either by a *very* short α-helix in case of the residues $(23, 33)$ and $(69, 90)$, or by a *very* short β-strand in case of residues $(51, 61)$.

We have performed a statistical analysis on the occurrence of our seven solitons in all PDB proteins. In Table 2 we list the number of matches that each of these solitons has when we search PDB for configurations that deviate from the given soliton by an overall root mean square distance (RMSD) which is less than 0.5 Å. We have chosen this cut-off value since it is representative of the Debye-Waller fluctuation distance in the experimental 1LMB data; see Figure 6. The interesting observation is that the second soliton of 1LMB that precedes

TABLE II: The (minimal) soliton sites that we have used in searching for matching structures in PDB, together with the number of matches. The search is limited to those x-ray structures that have a resolution better than 2.0 Å and a match is a configuration that deviates less than 0.5 Å in total root mean square distance (RMSD) from the soliton.

Soliton	1	2	3	4
Sites	(20,28)	(27,36)	(36,46)	(50,58)
Matches	9601	4	810	159

Soliton	5	6	7
Sites	(55,63)	(66,75)	(74,82)
Matches	1552	1342	406

the DNA recognition helix, is *unique* to the λ-repressor protein. The *only* matching structures are located in the different PDB entries of the λ-repressor protein itself. We also note that for this soliton the B-factors are slightly higher than for any of the other six solitons along the 1LMB backbone.

Finally, by following [12] we have combined the seven solitons together to describe the sites 8-90 (PDB indexing) of the entire 1LMB backbone. The result is shown in Figure 7. The overall RMSD accuracy that we get in

FIG. 7: (Color online) The PDB backbone configuration of 1LMB (dark red) and its multi-soliton (light green) approximation, for sites 8-90. The total RMSD distance is 0.45 Å.

this manner is 0.45 Å. This could be further improved by adjusting the parameters while combining the solitons. But as such, the accuracy displayed by the configuration in Figure 7 is below the experimental B-factors. Consequently any improvement is senseless, in light of the quality of the experimental data.

C: Soliton solution and 1LMB

We proceed to construct the parameters corresponding to the 1LMB backbone, for *both* the C_α-atoms and the side-chain C_β-atoms, using the energy function (35). We first solve (28) to obtain the bond angle *i.e.* κ_i-profile. We then construct the backbone torsion angles τ_i and the side-chain (θ_i, φ_i) angles from (23), (31), (34). We construct a *single* solution of (28), that describes the entire chain.

For simplicity of construction, we restrict the values of κ_i into the fundamental domain $\kappa_i \in [0, \pi]$. As in the case of the Ansatz (29), a higher precision could be obtained by allowing κ_i to take values beyond the fundamental domain. But with (35), it turns out that we can reach the B-factor accuracy by constructing the solution of (28), using the fundamental domain only. This is because the solution of (28) is even better suited for modeling proteins than the Ansatz (29).

To construct the parameters that describe the backbone κ_i profile of 1LMB, we use the iterative learning algorithm of [11]. It determines the parameters by locating the fixed point of

$$\kappa_i^{(n+1)} = \kappa_i^{(n)} - \epsilon \left\{ \kappa_i^{(n)} V'[\kappa_i^{(n)}] - (\kappa_{i+1}^{(n)} - 2\kappa_i^{(n)} + \kappa_{i-1}^{(n)}) \right\} \tag{40}$$

that minimize the RMSD distance to 1LMB. Here $\{\kappa_i^{(n)}\}_{i\in N}$ denotes the n^{th} iteration of an initial configuration $\{\kappa_i^{(0)}\}_{i\in N}$ and ϵ is some sufficiently small but otherwise arbitrary numerical constant. We select $\epsilon = 0.01$. A fixed point of (40) clearly satisfies the DNLS equation (28). Following [11] we utilize step-functions to construct an initial configuration for κ_i. The ensuing initial profile of $\kappa_i^{(0)}$ is chosen to have the same overall profile as the properly gauge transformed 1LMB that we display in Figure 4 right hand side column. A Monte Carlo routine is set up to determine the parameters. For this we have developed a package that we call *Propro* [24]. It implements our parameter learning algorithm for a given protein structure, largely automatizing the entire process.

In Figure 8 we show a *Propro* screen capture of the κ_i profile that describes the final multi-soliton solution that yields the shortest overall RMSD distance between the solution to (40) and the 1LMB structure, for the backbone C_α carbons. In Figure 9 we show the corresponding τ_i profile, computed from (23). The backbone C_α RMSD distance between out multi-soliton solution and 1LMB is 0.52 Å. This is slightly larger that what we obtained with the Ansatz (29). But this time we have not extended the values of κ_i outside of the range $\kappa_i \in [-\pi, \pi]$. In Figure 10 we display the distance between the 1LMB and the soliton solution to (28), (23) for the individual C_α atoms. For the most part the distance between our multi-soliton solution and 1LMB is below the Debye-Waller fluctuation distance. The only real exception is located at the site 31, where the distance between 1LMB C_α carbon and the soliton solution is close to 1.6 Å. This site is lo-

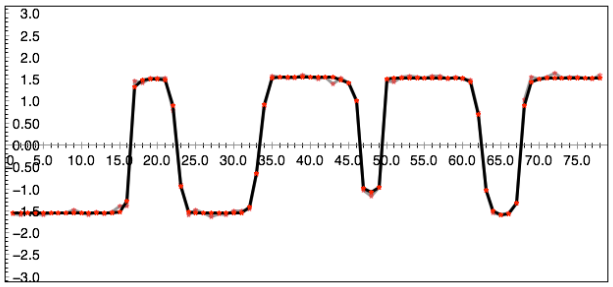

FIG. 8: (Color online) The multi-soliton solution to (40) (line) and the PDB values of 1LMB (dots) along the C_α backbone, for the backbone bond angles

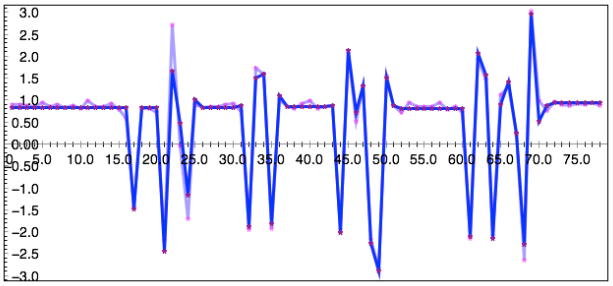

FIG. 9: (Color online) The profile of torsion angle computed from (23) (line) and the corresponding PDB values of 1LMB (red dots) along the C_α backbone.

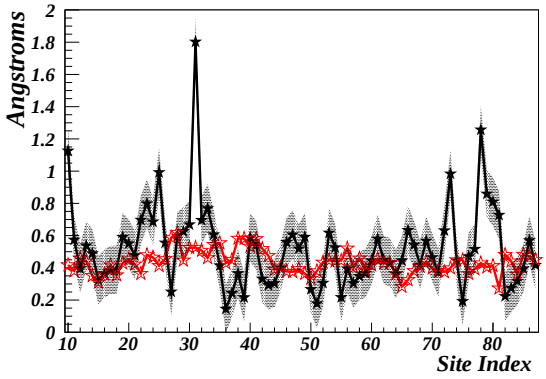

FIG. 10: (Color online) The distance between the 1LMB backbone and our multi-soliton configuration, constructed by solving (28) (black line). The grey shaded area around the black line describes the estimated 0.15 Å zero point fluctuation distance around the multi-soliton solution. The grey (red) line describes the experimentally measured Debye-Waller fluctuation distance

cated at the second soliton. We recall that this soliton is unique for 1LMB (see Table II) and that it was also singled out by the Ansatz (29). Our analysis indicates that something takes place at this soliton that warrants a more careful experimental analysis. The properties of this soliton might have a rôle in the transition from lysogenic to lytic state.

We proceed to extend our multi-soliton to describe both the positions of the C_α carbons, and the directions of the C_β carbons. For this we use (31) and (34). The final configuration has a combined C_α - C_β distance of 0.6 Å to 1LMB. In Figure 11 we compare the corresponding structures, the 1LMB backbone is translucent.

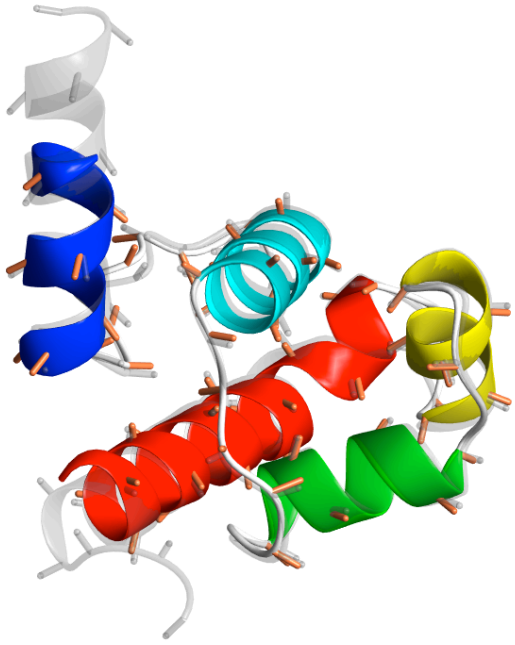

FIG. 11: (Color online) Comparison between the C_α - C_β multi-soliton solution (opaque, with colors online) and the 1LMB structure (translucent, grey online). The RMSD distance is 0.6 Å.

In Figure 12 we have a close-up of the region around PDB site 31, where the difference between the multi-soliton solution and the 1LMB configuration is largest.

Finally, in Table III we list the parameters in (35)-(39), for all the seven solitons.

IV: COLLAPSE STUDIES OF 1LMB

In the backbone energy (36), (37) we have retained only those variables that are relevant to our description of the C_α geometry. The derivation of (36), (37) is based on the general concept of universality [14]-[17], in combination with the requirement that the energy must be independent of the coordinate frame where it is computed. Consequently, *by construction*, our energy function correctly describes the leading order long distance contribution to *any* energy function that is grounded on more detailed atomic level considerations. *All* the variables and interactions that are less relevant for the description of the C_α geometry, are accounted for effectively through the functional form and the parameter values of the individual contributions to (36), (37).

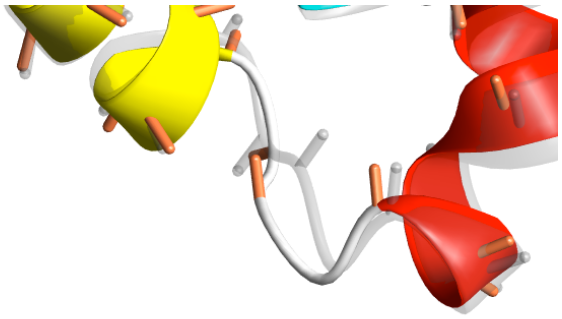

FIG. 12: (Color online) Close-up of Figure 11 around site 31, the location of the second soliton. Multi-soliton is opaque, the PDB structure of 1LMB is translucent.

We shall try and approach protein dynamics in the same universal manner. We average over all very short time scale oscillations, vibrations and other tiny fluctuations in the positions individual atoms that are basically irrelevant to the way how the folding progresses over those time scales that are biologically relevant. The general concept of universality [14]-[17] proposes us to adopt a Markovian Monte Carlo time evolution with the following universal, coarse grained heat bath probability distribution [25], [26], [27]

$$\mathcal{P} = \frac{x}{1+x} \qquad \text{with} \qquad x = \exp\{-\frac{\Delta E}{kT}\} \tag{41}$$

Here ΔE is the energy difference between consecutive MC time steps, that we compute from (36), (37). We select the numerical value of the temperature factor kT so that the model describes the appropriate phase. In [28] it has been shown that (36), (37) is capable of describing the three phases of polymers. At low values of kT we are in the phase of collapsed proteins. As the value of kT increases and reaches the Θ-point value, there is a transition to random coil phase. When the temperature reaches even higher values, there is a cross-over to self-avoiding random walk.

It turns out that in the collapsed phase, below the Θ-point temperature, the universal aspects of folding dynamics are quite independent of the numerical value of kT. For concreteness, we perform our simulations in the

collapsed using the value

$$kT = 10^{-15}$$

Note that the overall normalization of kT can always be changed by an overall normalization of the parameters in Table III.

We shall assume that during the folding process there are no re-arrangements in the backbone covalent bond structure, such as chain crossings. For this we introduce a self-avoidance condition that keeps the distance between any two backbone C_α atoms at least as large as the length of a typical van der Waals radius which is around $\sim$1.3 Ångström.

Note that we do not propose that (41) is capable of describing the atomic level dynamics of the folding process. Such minuscule details are highly sensitive to the initial atomic configuration. A detailed knowledge of the time evolution of a particular atom during the collapse can hardly have any practical relevance for the underlying physical principles and phenomena. Thus, for the purpose of conceptually understanding the temporal evolution of a protein towards its native conformation, the dynamics described by (41) is sufficient. We argue that the combination of (35), (41) correctly captures the universal statistical aspects of protein collapse over the biologically relevant temporal and spatial scales.

A: Antiferromagnetism and folding nuclei

In our approach, proteins have a modular structure. A folded protein is built by combining together solitons of the discrete non-linear Schrödinger equation (28), one after another. From the point of view of the energy function (36), (37), a uniform helical configuration is one with

$$\kappa_i \approx m$$

and with the value of τ_i computed from (23) this is a ground state of the energy. In particular, a straight linear rod is a special case, it is a ground state when $m = 0$.

As usual, the physical principles that give rise to protein folding are best analyzed in the absence of other processes and interfering agents. For this reason, in the present sub-section, we study the soliton formation along a helical backbone, by inspecting how the folded configuration builds from the ground state of the energy. Consequently we use a straight helical structure as our initial configuration. In the case of the λ-repressor protein we are particularly interested in the formation of the third, DNA binding soliton.

The parameter values in Table III are uniform over each putative super-secondary structure. In particular, there is no information on the loop locations. In a protein, the placement of a loop often correlates with the position of certain amino acids, such as proline and glycine, that act as folding nuclei. In order to model a folding nucleus we introduce a transient parameter σ that sends the first term in (36) into

$$-\sum_{i=1}^{N-1} 2\,\kappa_{i+1}\kappa_i \;\rightarrow\; \sum_{i=1}^{N-1} (2\sigma - 1)\cdot 2\,\kappa_{i+1}\kappa_i \qquad (42)$$

Initially, we set $\sigma = 0$ for all i except at the links $(i, i+1)$ along which the putative soliton centers are located. At these links we start with $\sigma = +1$. This corresponds to an anti-ferromagnetic *i.e.* repulsive nearest neighbor interaction between the ensuing two sites. During the early stages of the simulation, we decrease the values of σ so that after some number of steps we reach the uniform final value $\sigma = 0$ for all links along the entire chain.

For each super-secondary structure the value of m in Table III determines the regular secondary structure which is located either before or after the corresponding loop, and the value of the parameter q in (36) determines the propensity of this structure to form. The stability of α helices and β strands is due to hydrogen bonds that form during the collapse, and consequently the value of q can be interpreted as a measure of the strength of hydrogen bond interactions. In our simulations we wish to start from an initial configuration with no initial hydrogen bonds. We conform to this by setting all $q = 0$ initially. We then switch on the hydrogen bond interactions by increasing the values of the parameters q to those given in Table III. We find that this stabilizes the regular secondary structures.

We wish to investigate the effects that the temporal ordering of loop formation has on folding and on misfolding. For this we compare different orderings in removing σ and in switching on the values of the q.

In the case of the lysogeny maintaining λ-repressor protein we have considered various scenarios to conclude that there is the following general pattern: The seven solitons that we display in Figure 4 (right column) tend to form as pairs, with (2,3), (4,5) and (6,7) each a soliton-soliton pair. The first soliton is also made with a pair. But after the formation, the pair of this soliton moves away and disappears through the N-terminal of the backbone. The alternative, where the first and second soliton form as a pair seems to give rise to a misfolded state that furthermore appears to be unstructured. We now describe in detail two generic examples that illustrate this general pattern:

First example: In our first generic example we start from a uniform, straight helical structure. In the Figure 13a we show the initial κ_i profile. We note that there is substantial latitude in choosing the initial values of κ_i and τ_i. To begin with, we also set all $q = 0$ so that there are no hydrogen bonds to stabilize the helical structure. All the remaining parameters have the values that are listed in Table III. We introduce an antiferromagnetic coupling $\sigma = +1$ at links $(i, i+1)$ with

$i = 16, 23, 33, 46, 49, 63, 67$. This models the folding nuclei at the putative positions of the centers of the solitons. During the first part of the simulation, say during the first $2,000,000$ Monte Carlo steps, we adiabatically remove the folding nuclei by linearly decreasing the values of σ until we reach the final ferromagnetic values $\sigma = 0$ at all sites. At the same time we turn on the hydrogen bonds, by increasing the values of the couplings q from zero to the values given in Table III. After the first $2,000,000$ steps all parameters along the backbone then have the values shown in Table III.

We remind that according to our observations there is a lot of latitude in details of the procedure.

We find that the last four solitons form as the pairs (4,5) and (6,7). At the putative location of the third soliton, we also observe the formation of a soliton-soliton pair. But the soliton which is located closer to the N-terminal moves towards left as shown in Figure 13, and becomes anchored at the site 23 where it forms the second soliton of 1LMB. A new soliton pair appears at the putative location of the first soliton, one of these solitons disappears through the N-terminal and the entire backbone stabilizes rapidly into the correct native state. See Figure 13.

Second example: In this example we simulate a scenario where the first pair is formed before the third soliton. The initial configuration is a folded structure, with the fully formed soliton pairs (1,2), (4,5) and (6,7) *i.e.* these solitons have the same (κ_i, τ_i) values as the 1LMB. But the DNA binding third soliton is absent and instead there is a helix extending from the second to the fourth soliton, see Figure 14 and 15. The simulation starts with an antiferromagnetic $\sigma = 1$ at link $i = 34$, and with no hydrogen bond interactions *i.e.* $q = 0$ between second and fourth solitons. There is then an initial production of a soliton-soliton pair as seen in Figure 15b and we find two possibilities:

If the hydrogen bonds form slowly *i.e.* q grows to its final value slowly in comparison to the removal of σ, we arrive at a final fold where there is an additional soliton (loop) around site 39. We display the bond angle profile in Figure 14c. The protein is now misfolded. On the other hand, if the hydrogen bonds are formed more rapidly, we arrive at a configuration where the third and fourth soliton annihilate each other. No soliton is formed, but due to steric restraints there is a slightly irregular helical region between sites 31 and 45. The final configuration is shown in Figure 15.

We conclude that if the $(1, 2)$ pair forms before the third soliton, the protein becomes misfolded into a state with more than one conformational substrate. We propose that this could be confirmed experimentally. It might relate to the transition from the lysogenic to the lytic state in λ-phage.

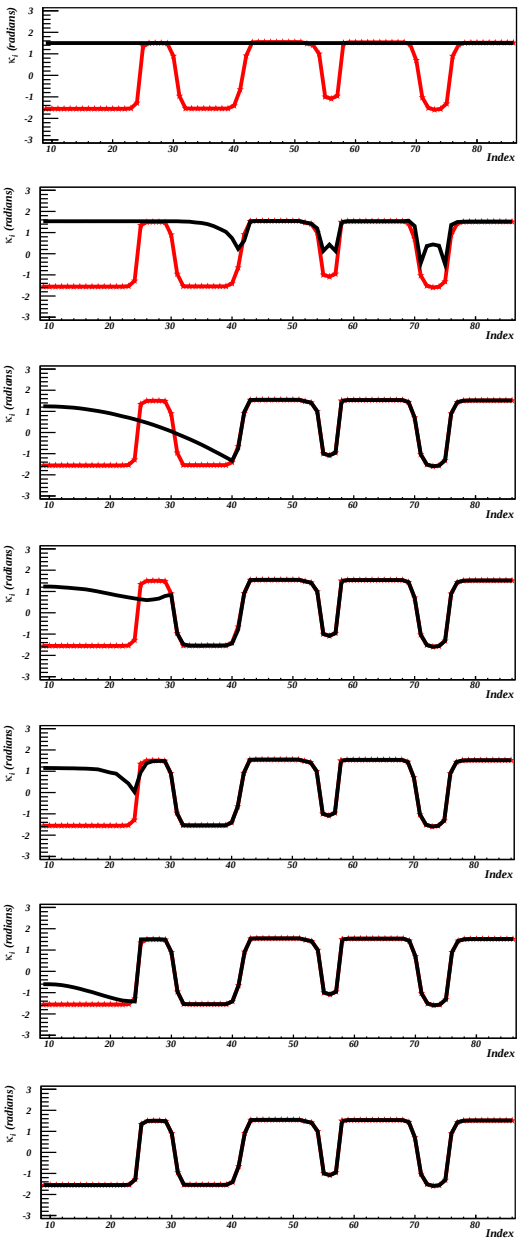

FIG. 13: (Color online) Time series of soliton (loop) formation along the 1LMB backbone in our simulation. The black line shows the time evolution of the solitons, the grey (red) line is the PDB profile. Time increases from top down. In the first Figure from the top we have the initial configuration, a straight helical structure. The solitons 4-7 form as the two soliton pairs (4,5) and (6,7). At the position of soliton 3, there is also a pair formation. But the pair of soliton 3 propagates along the chain towards the N-terminal, until it becomes anchored at site $i = 23$ where it forms the second soliton. A new soliton pair forms at site 16 and one of these two solitons moves and disappears through the N-terminal, leaving us with the first soliton and the correctly folded backbone

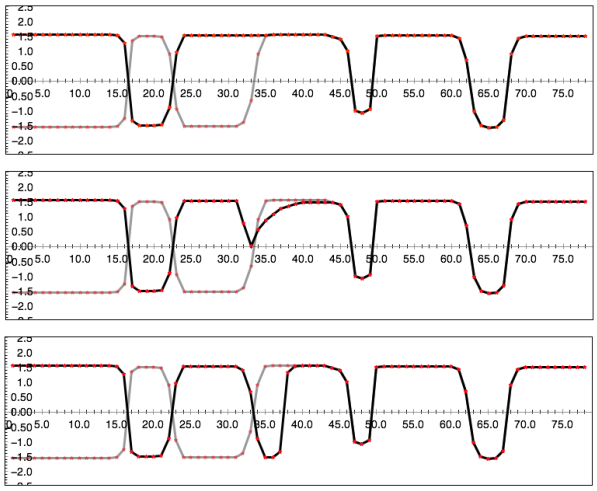

FIG. 14: (Color online) If hydrogen bonds form slowly, there is an extraneous soliton that disturbs binding between 1LMB and DNA, and probably no binding is possible. Grey line is the 1LMB profile, and black with red dots is the simulated profile.

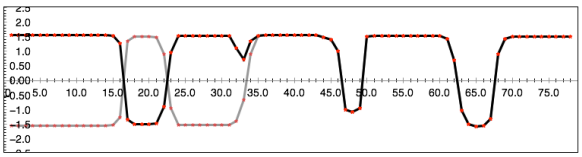

FIG. 15: (Color online) If the hydrogen bonds are formed early, the backbone enters in an unstructured state where the 3 soliton become mis-formed. Consequently there is no soliton that would dock the 1LMB with DNA. Grey line is the 1LMB profile, and black with red dots is the simulated profile.

B: Comparison of λ-repressor and CRO soliton structures

In Figure 16 we compare the first two solitons in 1LMB to the first loop in the CRO regulator protein that controls the transition to the lytic state. For the latter, we use the PDB structure with code 2OVG. In the CRO protein, the first loop is topologically more stable, in the sense that it consists of a single soliton. As such the loop in 2OVG is more stable than in 1LMB. For a plane curve, a single soliton can be made or deleted only by transporting it through one of the end points of the curve. On the other hand, a pair of solitons such as the one in the left hand side of Figure 16 is not topologically stable but can be more easily created or removed locally, by a saddle-node bifurcation that brings the two solitons together. This removes the corresponding loop by converting it (in this case) into a single long α-helix.

A comparison between the λ-repressor and CRO profiles in Figure 16 proposes the following mechanism for the lysogenic-lytic transition: Under lysogenic conditions where the λ-repressor protein prevails, the soliton pairs of the λ-repressor protein that are located immediately prior and after the DNA binding domain are relatively stable. But when there is a change in the environmental conditions that excites phonon fluctuations along the protein chain such as raise in temperature or UV radiation, or maybe an enzymatic action that remains to be identified, either of these soliton pairs can discharge by a saddle-node bifurcation. This bifurcation disturbs the structure of the immediately adjacent DNA binding motif to the extent that the protein looses its capability to maintain the lysogenic phase. Since each of the corresponding motifs in the CRO protein are topologically more stable single soliton configurations, they are much more insensitive to effects such as local phonon excitations due to UV radiation and thermal effects, and consequently the lytic phase can take over. we display the

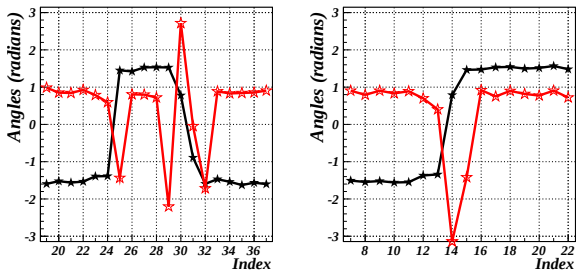

FIG. 16: (Color online) The resolved (κ_i, τ_i) spectrum for the first PDB helix-loop-helix of 1LMB (left) and the corresponding structure of in 2OVG (right). The bond angle κ is black, torsion angle τ is grey (red). The bond angle spectra reveal that in 1LMB the loop is a bound state of two close-by solitons while in 2OVG there is only one soliton profile.

first PDB helix-loop-helix motif and for comparison we display the corresponding structure in the CRO protein with PDB code 2OVG. In the case of λ-repressor the motif is clearly a bound state of two solitons while in the case of CRO we have a single isolated soliton.

C: Heating and cooling in 1LMB

We apply (41) to theoretically investigate what takes place in 1LMB when we heat it into the random coil phase, and then re-cool it back to the collapsed phase. According to Anfinsen [29] the protein should return to its original conformation.

We start from the multi-soliton configuration that describes the PDB structure 1LMB with parameter values given in Table III. Unlike in the previous subsection, during the entire heating and cooling cycle we now keep all

the parameter values intact. In particular, neither during the heating nor during the cooling do we introduce any transient antiferromagnetic parameter σ as in the previous subsection. Nor do we change the parameter values q during the present process. As a consequence the position of any soliton and the size of any helical structure becomes determined dynamically, without any explicit folding nuclei. Moreover, since the parameter values q do not change, the strength of the underlying hydrogen bond interactions remains constant during the simulations. Any deformation or adjustment in the helical structures will be entirely due to thermal fluctuations during the heating and cooling cycle.

We introduce the heat bath dynamics (41). The temperature kT is assumed to be globally determined, and the heating and cooling should proceed slowly over the biologically relevant time scales. This ensures that the entire protein structure is kept at an equal temperature value so that we can ignore any effects due to local temperature variations.

During heating and cooling cycle we follow the backbone evolution by computing the root-mean-square distance (RMSD) between the C_α coordinates $\mathbf{r}_{0i}$ of the 1LMB backbone and those of the multi-soliton configuration $\mathbf{r}_i$, as a function of the temperature

$$RMSD(T) \stackrel{def}{=} \sqrt{\frac{1}{N}\sum_i (\mathbf{r}_{i0} - \mathbf{r}_i)^2} \qquad (43)$$

We start from a low temperature value that corresponds to the collapsed phase. We again choose the numerical value

$$kT \;=\; 10^{-15} \qquad (44)$$

for the initial configuration. We adiabatically increase the temperature kT to some high value during 500,000 MC steps. We then keep the system at this high temperature during 1,000,000 MC steps, and finally cool it back to the original temperature value (44) during another 500.000 steps. The relatively large number of MC steps in our cycle at the high temperature ensures that the protein becomes fully thermalized to that temperature value.

We have made simulations with several hundreds of repeated heating and cooling cycles, always starting with (44) and heating to different high temperature values that are well above the Θ-point, where the protein enters the random walk phase. From Figure 17 we learn that as long as the high temperature value remains below

$$kT_{C1} \;\approx\; 10^{-4} \qquad (45)$$

the protein always returns back to its original shape upon cooling. But between values in the range

$$10^{-4} \approx kT_{C1} < kT < kT_{C2} \approx 10^{-2}$$

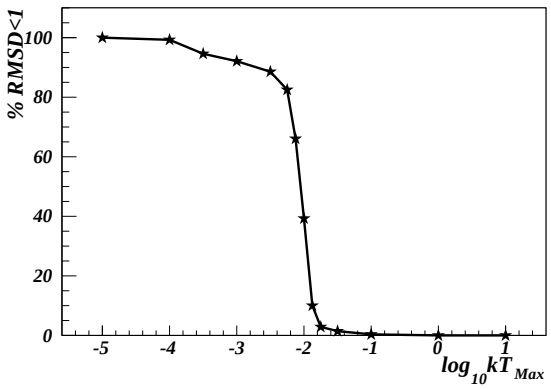

FIG. 17: At temperature values $\log kT < -4$ the protein returns to its original conformation after the heating and cooling cycle. Between $-4 < \log kT < -2$ a small fraction starts to misfiled. At $\log kT \approx -2$ there is a rapid transition, so that proteins that have been heated to higher temperatures become misfolded.

there is a small fraction of cycles at the end of which the protein becomes misfolded. Finally at around the temperature value kT_{C2} there is a rapid cross-over and if the heating temperature exceeds

$$T > T_{C2}$$

the final conformation is always misfolded. The misfolding is caused by deformation of one or more of the loops, often due to the wrong ordering in loop formation.

In Figure 18 we show as an example, how (43) evolves in average during a typical heating cycle that reaches very near the critical temperature value (45). We find that during the heating there is a rapid increase in the RMSD distance. This is due to the transition to the random coil phase. When the system is kept at the high temperature, there are substantial thermal fluctuations. The shaded region (blue online) denotes the one standard deviation extent of these fluctuations. During the cooling we have a collapse transition, and at the end the value of (43) becomes small with practically no fluctuations, indicating that the protein has returned to the original 1LMB like conformation. In Figure 19 we display a generic random coil structure that we observe during the heating phase. It has the look of a typical random coil with no regular, helical structure remaining. The structure is not static, but subject to very strong thermal fluctuations in its shape.

SUMMARY

In summary, we have investigated the structure and folding pathways of the lysogeny maintaining λ-repressor protein. Our approach is based on an effective energy

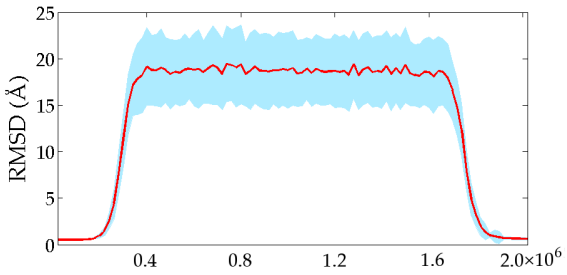

FIG. 18: (Color online) The evolution of the RMSD value (43) between the 1LMB and the heated structure during one cycle. At the high temperature value which is here $kT = 10^{-4}$, the RMSD is large and subject to large thermal fluctuations; the shaded area denotes the one standard deviation fluctuation regime in the values of (43). At the end of the cycle, the protein folds back to the conformation of 1LMB.

FIG. 19: A generic configuration during the high temperature regime in the cycle of Figure 18. The structure is an apparent random coil with no regular structure, and its detailed form is subject to strong thermal fluctuations.

function, that we have argued describes the small fluctuation limit of *any* atomic level energy function. Our justification of the energy function is entirely based on two very general physical principles: The concept of universality originally introduced in the context of phase transitions and critical phenomena, and the demand that any physical quantity must be independent of the coordinate system that is used for its description. Using these two universal physical principles, we have shown that the long wavelength fluctuation limit of the energy function for both the backbone C_α and side-chain C_β atoms is fully determined in an essentially unique fashion.

The energy function we have derived, computes the C_α and C_β conformation in terms of a soliton solution to a variant of the discrete non-linear Schrödinger equation as the modular component. The explicit form of the relevant dark soliton solution is not known to us, but we have found that an excellent approximation can be obtained by naively discretizing the known dark soliton of the continuum non-linear Schrödinger equation. We are able to re-construct the entire backbone of the λ-repressor protein with a precision that compares and even exceeds the precision of the experimental crystallographic structure with PDB code 1LMB, when we determine the precision in terms of the Debye-Waller B-factor fluctuation distance. The high precision of our theoretical description enables us to conclude that the second soliton solution that appears in our description of the 1LMB is *unique* to this protein, there are no other similar solitons in the entire Protein Data Bank. The remaining solitons including the DNA binding one are all ubiquitous in PDB. We have also investigated the corresponding soliton structure in the CRO protein that is responsible for the lytic phase, and found that this soliton appears more stable and is also commonly found in the PDB data. These observations suggest that the transition between the lysogenic and lytic life-cycles could somehow relate to the very exceptional structure of the second soliton in 1LMB.

We have extended our energy function to describe the collapse dynamics of 1LMB. In line with the construction of the energy function, we rely on the concept of universality to propose that at biologically relevant time scales the folding dynamics can be described in terms of Glauberian relaxation dynamics, with Markovian time evolution. In this way we have found that all solitons in 1LMB appear to form as soliton-soliton pair. In particular, the second soliton with its exceptional structure forms as a pair with the DNA binding third soliton; The pair of the first soliton flows away and disappears through the N-terminal of the protein. Moreover, if for some reason the formation of the second and third solitons is disrupted, for example if the second soliton forms by itself before the third soliton, we find that the third soliton can not form properly. It is either formed in combination of an extra soliton, or then the protein enters in an unstructured conformation. The choice between these two alternatives is made by the strength of the hydrogen bond formation.

A.N. thanks H. Frauenfelder and G. Petsko for communications and J. Åqvist for discussions.

* Electronic address: Andrei.Krokhotine@cern.ch
† Electronic address: martin.lundgren@gmail.com
‡ Electronic address: Antti.Niemi@physics.uu.se

[1] M. Ptashne, A Genetic Switch, Third Edition, Phage Lambda Revisited. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (2004)
[2] M.E. Gottesman and R.A. Weisberg, Microbiol. Mol. Biol. Rev. **68** 796 (2004)
[3] A. Amir, O. Kobiler, A. Rokney, A.B. Oppenheimer and J. Stavans, Mol. Syst. Biol. **3** 1744 (2007)
[4] B. Snijder and L. Pelkmans, Nature Reviews, Mol. Cell.

Biol. **12** 119 (2011)
[5] F. St-Pierre and D. Endy, Proc. Nat. Acad. Sci **105** 20705 (2008)
[6] T. Wigle and S. Singleton, Bioorg. Med. Chem. Lett. **17** 3249 (2007)
[7] M. Avlund, I. Dodd, S. Semsey K. Sneppen and S. Krishna, Journ. Virol. **83** 11416 (2009)
[8] H.M. Berman, K. Henrick, H. Nakamura, J.L. Markley, Nucl. Acids Res. **35** (Database issue) D301 (2007)
[9] U.H. Danielsson, M. Lundgren and A.J. Niemi, Phys. Rev. **E82** 021910 (2010)
[10] M. Chernodub, S. Hu and A.J. Niemi, Phys. Rev. **E82** 011916 (2010)
[11] N. Molkenthin, S. Hu and A.J. Niemi, Phys. Rev. Lett. **106** 078102 (2011)
[12] S. Hu, A. Krokhotin, A.J. Niemi and X.Peng, Phys. Rev. **E83** 041907 (2011)
[13] S. Hu, M. Lundgren and A.J. Niemi, arXiv:1102.5658 Biomolecules (q-bio.BM); Phys. Rev. **E** (to appear)
[14] B. Widom, J. Chem. Phys. **43** 3892 (1965)
[15] L.P. Kadanoff, Physics (Long Island City, NY) **2** 263 (1966)
[16] K.G. Wilson, Phys. Rev. **B4** 3174 (1971)
[17] M.E. Fisher, Rev. Mod. Phys. **46** 597 (1974)
[18] A.J. Niemi, Phys. Rev. **D67** 106004 (2003)
[19] M. Chernodub, L.D. Faddeev, A.J. Niemi, JHEP 0812:014 (2008)
[20] L.D. Faddeev, L.A. Takhtajan *Hamiltonian methods in the theory of solitons* (Springer Verlag, Berlin, 1987)
[21] A. Krokhotin, A.J. Niemi, X. Peng Phys. Rev. **E85** (2012) 031906
[22] M. Lundgren, A.J. Niemi, S. Fan, Phys. Rev. **E** (in print)
[23] M. Lundgren, A.J. Niemi, to appear
[24] Available from `http://folding-protein.org`
[25] R.J. Glauber, Journ. Math. Phys. **4** 294-207 (1963)
[26] A.B. Bortz, M.H. Kalos and J.L. Lebowitz, Journ. Comput. Phys. **17** 10-18 (1975)
[27] A. Krokhotin, M. Lundgren, A.J. Niemi, (to appear)
[28] M. Chernodub, M. Lundgren, A.J. Niemi, Phys. Rev. **E83** 011126 (2011)
[29] C.B. Anfinsen, Science **181** 223 (1973)

TABLE III: Our best parameter values for the multi-soliton solution that models 1LMB. Notice that in line with (26), (27), the values of m_1, m_2 imply that all the regular structures are α-helices except for the one separating solitons 4 and 5 which is a short β-strand. We also note that the overall scale of the parameters is fixed by the normalization of the first term in (36) which we have chosen for convenience.

Soliton	q_1	q_2	m_1	m_2
1	1.12091	1.87372	1.55737	1.50013
2	0.357906	9.69166	1.65666	1.54119
3	0.260909	6.14144	1.68182	1.54676
4	0.684119	4.75578	1.47243	1.09234
5	5.15882	6.77828	1.07066	1.53464
6	0.314503	0.38534	1.66164	1.6235
7	0.947322	0.624884	1.58568	1.51678

Soliton	a	b	$c/2$	$d/2$
1	-1.0	-32357500	47.5340	438750
2	-1.0	-1.86689	0.0413976	0.00101952
3	-1.0	-9.26604	0.121919	0.0271569
4	-1.0	-7.51371	0.013751	0.0148377
5	-1.0	-25.0125	0.251803	0.597751
6	-1.0	-23.9299	0.0181312	0.0410994
7	-1.0	7.9809	0.000215793	0.037091

Soliton	a_θ	$b_\theta \cdot 10^{-12}$	c_θ	d_θ
1	1.24035	-475.728	1.0	$5.44473 \cdot 10^{-10}$
2	2.74724	-198463	1.0	0.42667
3	1.32293	$-5.021439 \cdot 10^{8}$	1.0	$3.74966 \cdot 10^{-6}$
4	1.34998	-81.1641	1.0	$6.59854 \cdot 10^{-9}$
5	1.38447	-6629519	1.0	$5.20625 \cdot 10^{-7}$
6	1.38293	-6.28532	1.0	$1.34992 \cdot 10^{-5}$
7	1.23869	-177.296	1.0	$7.6271 \cdot 10^{-7}$

Soliton	a_φ	b_φ	c_φ	$d_\varphi \cdot 10^{-10}$
1	0.971374	$-6.7206 \cdot 10^{-10}$	1.0	7211.16
2	0.813254	$2.3582 \cdot 10^{-7}$	1.0	1481.99
3	0.771272	$2.74 \cdot 10^{-6}$	1.0	-51.6611
4	0.616865	$1.2158 \cdot 10^{-11}$	1.0	5478
5	0.89315	$3.87 \cdot 10^{-9}$	1.0	1039.56
6	0.545154	0.184472	1.0	1.2685
7	0.988183	$3.12459 \cdot 10^{-9}$	1.0	2.09412